Expanding duplication of the testis PHD Finger Protein 7 (PHF7) gene in the chicken genome

Sophie Fouchécourt, Valérie Fillon, Christelle Marrauld, Caroline Callot, Sarah Ronsin, Floriane Picolo, Cécile Douet, Benoit Piegu, Philippe Monget

To cite this version:

Sophie Fouchécourt, Valérie Fillon, Christelle Marrauld, Caroline Callot, Sarah Ronsin, et al.. Expanding duplication of the testis PHD Finger Protein 7 (PHF7) gene in the chicken genome. Genomics, 2022, 114 (4), pp.1-9. 10.1016/j.ygeno.2022.110411 . hal-03746220

HAL Id: hal-03746220

https://hal.inrae.fr/hal-03746220

Submitted on 22 Jul 2024

Expanding duplication of the testis PHD Finger Protein 7 (*PHF7*) gene in the chicken genome

Running title: The *PHF7* gene expansion in the chicken genome

Sophie Fouchécourt≠,[1], Valérie Fillon[2], Christelle Marrauld[2], Caroline Callot[3], Sarah Ronsin[1], Floriane Picolo[1], Cécile Douet[1], Benoit Piégu[1], Philippe Monget[1]

1) CNRS, IFCE, INRAE, Université de Tours, PRC, F-37380, Nouzilly, France
2) GenPhySE, Université de Toulouse, INRAE, ENVT, F-31326, Castanet Tolosan, France
3) CNRGV – Plant Genomic Center INRAE F-31326, Castanet Tolosan, France

≠Corresponding author: sophie.fouchecourt@inrae.fr

Competing interests: The authors declare that they have no competing interests.

Funding: This work was supported by grants from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement N°677353 (IMAGE), the Région Centre Val de Loire (FERTILMALE: n° 2017 119977), the French National Infrastructure of Research CRB Anim funded by 'Investissements d'avenir', INRAE (Institut National de la Recherche pour l'Agriculture, l'alimentation et l'Environnement). F. Picolo received a fellowship from CRB Anim and S. Ronsin received a fellowship from FERTILMALE.

Authors' contributions: SF, performed chicken gene analyses, polymerase chain reaction, generated all figures and drafted the manuscript; VF and CM, performed fluorescence *in situ* hybridisation; CC, in charge of PacBio sequencing; SR and BP, contributed to genomic annotation; FL, initial tracking of chicken orthologues; CD, contributed to genomic polymerase chain reaction; PM, designed the study, supervised the project and revised the manuscript. All authors read and approved the final manuscript.

ABSTRACT

Gene duplications increase genetic and phenotypic diversity and occur in complex genomic regions that are still difficult to sequence and assemble. PHD Finger Protein 7 (PHF7) acts during spermiogenesis for histone-to-histone protamine exchange and is a determinant of male fertility in *Drosophila* and the mouse. We aimed to explore and characterise in the chicken genome the expanding family of the numerous orthologues of the unique mouse *Phf7* gene (highly expressed in the testis), observing the fact that this information is unclear and/or variable according to the versions of databases. We validated nine primer pairs by *in silico* PCR for their use in screening the chicken bacterial artificial chromosome (BAC) library to produce BAC-derived probes to detect and localise *PHF7*-like loci by fluorescence *in situ* hybridisation (FISH). We selected nine BAC that highlighted nine chromosomal regions for a total of 10 distinct *PHF7*-like loci on five *Gallus gallus* chromosomes: Chr1 (three loci), Chr2 (two loci), Chr12 (one locus), Chr19 (one locus) and ChrZ (three loci). We sequenced the corresponding BAC by using high-performance PacBio technology. After assembly, we performed annotation with the FGENESH program: there were a total of 116 peptides, including 39 PHF7-like proteins identified by BLASTP. These proteins share a common exon-intron core structure of 8–11 exons. Phylogeny revealed that the duplications occurred first between chromosomal regions and then inside each region. There are other duplicated genes in the identified BAC sequences, suggesting that these genomic regions exhibit a high rate of tandem duplication. We showed that the *PHF7* gene, which is highly expressed in the rooster testis, is a highly duplicated gene family in the chicken genome, and this phenomenon probably concerns other bird species.

Key words: chicken genome, gene duplication, evolution, testis

INTRODUCTION

Gene families correspond to genes clustered by sequence similarity. The members often exhibit similar functions and evolve in a dynamic context of genomic rearrangements, including duplications within a single genome [1]. Moreover, it is well known that copy number variation (CNV) in loci with numerous paralogous genes has the potential to increase phenotypic diversity [2]. Indeed, CNV has a central role in explaining innovations across phyla, including the emergence of novel functions. We have analysed several duplicated gene families involved in reproduction in the mouse such as *Oogenesins* [3, 4] and *Nlrp* (Nucleotide-binding oligomerization domain, Leucine rich Repeat and Pyrin domain containing Proteins) [5, 6]. In the chicken genome, several genes are duplicated, most of which are simple tandem duplications, such as *DEFENSIN*6/7 [7]. However, some genes present several paralogues – for example, 10 for the Toll-like receptor gene family [8], 16 for the *TCRb* (T cell receptor beta) locus [9] and 23 for Free Fatty Acid Receptor-2 (*FFAR2/GPR43*) characterised previously in our laboratory [10].

Numerous genes are required for the formation of fertile spermatozoa (at least 2,000 different genes in humans), in particular genes involved in the successive stages of spermatogenesis (germ cell differentiation) [11]. Despite a huge number of studies, the causes of fertility defects (genetic or environmental perturbations) are not well understood, suggesting that there are still regulatory pathways to decipher, especially in species like birds in which testicular data are scarce compared with mammals. At the cellular and physiological levels, spermatogenesis represents a relatively well-conserved process among phylogenetically distant animal species [12, 13].

In a previous study published in 2019 [14], we were interested in identifying testis genes conserved between invertebrates and vertebrates and exhibiting high relative mRNA expression in the testis of vertebrates, with a focus on chicken species. Indeed, this study highlighted a substantial list of uncharacterised genes for testis function in vertebrates, in particular in the chicken. Among the chicken testis-specific genes highlighted in this previous study, PHD Finger Protein 7 (*Phf7*) is indispensable for mouse and *Drosophila* male fertility [15, 16], encodes an actor in histone-to-protamine exchange during spermiogenesis and is highly expressed in the mammalian and chicken testis [14, 17, 18]. We have been intrigued by the numerous bird orthologues of *Phf7* gene [14], as shown in its phylogenetic tree of EnsEMBL database (63 paralogues in EnsEMBL release 88 [March 2017]), whereas the gene is present in a single copy in mammals. Wang et al. [18] described at least two chromosomal loci in the chicken genome (with no precision about the number of paralogues). To our knowledge, such a high number of gene duplicates (63) in chicken has not been described in the literature. Moreover and of note, we observed that EnsEMBL chicken *PHF7* paralogues in releases following 88 have fluctuated (68, then 0, currently four paralogues), suggesting that these predictions are incomplete and/or unreliable. Indeed, while mouse and human genomes are high quality, this is not the case for more recently sequenced genomes, in particular the chicken genome, because it contains numerous GC-rich regions. It is known that gene duplications, especially when they are multiple and in tandem, correspond to complex genomic regions that are difficult to sequence, assemble and annotate. The scientific community has made an effort to fill these lacunae due to such technological and bioinformatic challenges. Thus, our objective was to explore and characterise the existence of this predicted expanding family of numerous *PHF7* orthologues in the chicken genome, using the high-performance PacBio technology [19].

RESULTS

Statement of the existence of an expanding family of *PHF7* genes in the chicken genome

In our 2019 study using data from the EnsEMBL 88 database released in March 2017, we noted that the mouse *Phf7* gene (present in a single copy in mammalian genomes) exhibited 111 avian orthologs for only five bird species, as schematised in the trees in Supplemental Fig 1(A/B) (see also Supplemental Table 1), with 63 homologs in the chicken, and between 2 and 21 in the four other birds (flycatcher: 2; duck: 6; zebra finch: 19; turkey: 21). As seen in the tree in Supplemental Fig 1B, the 63 predicted chicken genes were phylogenetically regrouped in nine subtrees. We arbitrarily named these subtrees: groups (Gr) A to I, for better clarification and further investigation. According to EnsEMBL 88, these genes are located on nine distinct loci on five chromosomes: Chr19 (GrA), ChrZ (GrB and GrC), Chr1 (GrD, GrE and GrI), Chr2 (GrF and GrH) and Chr12 (GrG). Surprisingly, we have noticed that the number of chicken *PHF7* genes annotated in EnsEMBL has fluctuated (from 0 to 68 genes) according to database versions (Supplemental Table 2), with only four orthologues described in the current version (release 105, [December 2021]) (Supplemental Table 2, column K). Thus, we decided to evaluate precisely the number of *PHF7* copies in the chicken genome by *in silico* PCR and genomic sequencing of bacteria artificial chromosomes (BAC) corresponding to genomic regions identified by fluorescence *in situ* hybridisation (FISH).

Characterisation of the chicken *PHF7* loci by *in silico* PCR

Primers designed in our previous study [14] and corresponding to the nine GrA to GrI *PHF7* subtrees/groups described above are listed in Supplemental Table 3 (column B). We first verified that these primers amplified unique amplicons with the expected sizes (Supplemental Table 3, column C) when used in genomic PCR with chicken DNA (Supplemental Fig 2). We then used these primer pairs in *in silico* PCR (UCSC database). They highlighted nine distinct genomic loci described in Supplemental Table 3 (column D). These loci are on the same chromosome for each respective group of genes as those described in the EnsEMBL 88 release: two loci on Chr19 corresponding to GrA; two loci on ChrZ corresponding to GrB and GrC; three distinct loci on Chr1 corresponding to GrD, GrE, and GrI; two loci on Chr2 corresponding to GrF and GrH; and one locus on Chr12 corresponding to GrG. Thus, these results confirm the existence of at least nine loci on five chromosomes with *PHF7* genes in the chicken genome as suggested by EnsEMBL version 88, in contrast with the data available in the current EnsEMBL version 105 that describes four loci on two chromosomes (Chr12 and Chr1, see Supplemental Table 2, column L). This allowed us to validate the relevance of the primer pairs for their use in BAC screening further followed by FISH.

Chromosomal localisation of chicken *PHF7* loci by FISH and comparison with *in silico* PCR localisation

For the nine groups, BAC clones were selected by PCR with primer pairs characterised above and as described in Materials and methods. We then used these specific BAC clones, listed in Table 1, in FISH to map the nine groups of genes (GrA to GrI), after we validated the nucleotide sequence of each probe by sequencing (sequences obtained are in Supplemental Table 3, column E) and submitted them to BLASTN analysis (Supplemental Table 3, column F). The GrA and GrB sequence probes were associated with an accession number with the name 'PHD finger protein 7_like' (GrA: NC_006106.5; GrB NC_006127.5), this is why we use '*PHF7*-like' for further designation in the text. The coordinates in BLASTN results were coherent with the loci delivered by *in silico* PCR (Supplemental Table 3, column D). The FISH localisations are shown in Fig 1A (pictures) and drawn in Fig 1B (chromosomal schemes), and listed in Table 1 (with their measures) and in Supplemental Table 3. FISH localisations specified a unique chromosomal region with *PHF7*-like genes for each BAC, except GrG that brightened two loci, one on Chr12 and one on ChrZ (this latter was not predicted by *in silico* PCR). GrD and GrE brightened the same chromosomal region on Chr1 (p26). There were a total of nine hybridisation signals dispatched on **five *G. gallus* chromosomes**: **Chr1** (two signal), **ChrZ** (three signal), **Chr2** (two signal), **Chr12** (one signal) and **Chr19** (one signal).

Sequencing, assembling and gene annotation of BAC clones

To further characterise the loci identified on the chicken genome by FISH, we sequenced the nine BAC clones targeted by the *PHF7* screen (sequences are available in Genbank[1]). The sequences contained between 87,395 bases for the smallest BAC (E) to 177,594 for the largest (I) (Table 2). For each of the nine nucleotide sequences, we performed BLASTN by using the NCBI database (GCRg7w -white Leghorn race- version 106). The BLASTN match result was unique for each BAC, except for BAC G that first matched with Chr12 for almost its entire length and then with ChrZ for only 19% of its length (see Supplemental Table 3, columns H and I). As excepted, the loci defined by these coordinates included the loci identified by *in silico* PCR for all BAC. We noted that BAC E and BAC D, which both hybridised on GGA1p26 in FISH as stated above, are 2.5 Mb apart and thus define two distinct *PHF7*-like loci on Chr1. Finally, these results confirm the chromosomal localisation and coordinates

[1] GenBank accession numbers: ON022098 (BAC A), ON022099 (BAC B), ON022100 (BAC C), ON022101 (BAC D), ON022102 (BAC E), ON022103 (BAC F), ON022104 (BAC G), ON022105 (BAC H), ON022106 (BAC I)

described above for the nine BAC as well as the existence of a tenth locus on ChrZ (GGAZ p21-22 corresponding to BAC G) that was not predicted by *in silico* PCR.

We performed gene annotation of the nine sequences by using the FGENESH program (in Softberry) as described in the Materials and methods. We identified 116 predicted genes dispatched on the nine BAC, from 4 for BAC F to 28 for BAC I (Table 2). Details of each BAC annotation FGENESH outputs are listed in Supplemental Table 4. We submitted the corresponding 116 peptides to BLASTP to determine functional homology. We retrieved a name and accession number for each of the 116 peptides (Supplemental Table 5). The annotation of each BAC is illustrated by gene maps in Fig 2 (BAC I) and Supplemental Fig 3 (BAC A to H). There were a total of 39 PHF7-like proteins (in yellow) from 1 in BAC C, BAC F and BAC H to 10 in BAC I (see Supplemental Table 5 and Table 2). The BLASTP results for these 39 PHF7-like proteins (sequences in Supplemental Table 6) corresponded to an E-value equal or very close to zero (using ‘by default’ parameters) (Supplemental Table 5). Based on using ESPript for obtention of a consensus sequence, their sequence similarity was 30%–55%, with a consensus sequence of 265 amino acids (Supplemental Data; see also the phylogenetic tree Fig 3).

Genomic organisation and phylogenetic link of the PHF7-like protein family

<u>Exon-intron structure</u>

We retrieved the exon sequence from each *PHF7*-like gene from FGENESH outputs and then used BLASTN to compare the sequences against one another. Comparative analysis of the exon/intron structure of the 39 *PHF7*-like genes revealed a frequent common ‘core-structure’ containing from 8 (1 gene) to 11 exons (11 genes), with the most prevalent configuration (16 genes) exhibiting 9 exons (Supplemental Table 7 and Supplemental Fig 4). This common ‘core-structure’ exhibits various changes: exon deletion (for example, in BAC B gene 20, exon 8 is deleted compared with other neighbour genes); reverse duplication on the other strand (for example, genes 3 and 4 in BAC G); split of an exon, that is, one exon gives two exons (for example, genes 3 and 6 in BAC E); and the presence of specific exons (genes 7 and 8 in BAC E). Finally and intriguingly, four genes (genes 4 and 5 in BAC E; gene 1 in BAC F; gene 5 in BAC H) were predicted to be longer because of several duplications of the ‘core-structure’. Gene 5 in BAC H was particularly intriguing (70 exons: 7 repetitions of 10 exon structures) and was also predicted by the annotation obtained with two other classical programs, Augustus and GENESCAN (not shown). On the contrary, the three other genes (predicted with 30, 29, and 28 exons, respectively for genes 4 and 5 in BAC E, and gene 1 in

BAC F) were split in 'normal-sized' *PHF7*-like genes with these programs (see the Discussion for more details).

Phylogeny

We aligned the predicted protein sequences to construct a phylogenetic tree that exhibited reliable bootstrap values (Fig 3). The phylogenetic link designates the peptides in BAC A as the sequences closest to the common ancestor. The phylogeny showed that the duplications occurred first between chromosomal regions and then inside each region. For example, on ChrZ, B and C loci are around 4 Mb apart and the tree suggested that one B peptide probably duplicated in the unique C peptide before its own multiple tandem duplications. The same scenario is observed for D and E peptides, which are separated by around 2 Mb on Chr1.

Other duplicate gene families in the *PHF7*-like loci

In each BAC, we noticed several groups of genes for which BLASTP results gave the same name and/or Genbank ID (Supplemental Table 5, Fig 2 and Supplemental Fig 3), suggesting that they are phylogenetically related and are members of the same family. There are 1) families with two or three members: ras GTPase-activating protein 1 in BAC B, Cadherin-18 in BAC F, scm-like with four MBT domains protein 2 in BAC D, testis-expressed protein 264, metabotropic glutamate receptor 2 in BAC G and centrosomal protein of 126 kDa in BAC I; or 2) families with many members: chemokine in BAC A (6 genes), reverse transcriptase dispatched in BAC B, BAC C and BAC I (16 genes); and translation initiation factor IF-2-like (9 genes) in BAC I. Interestingly, the genomic organisation of this latter BAC suggests tandem duplication of 'PHF7-like/translation initiation factor' genes. Overall, this confirms that these chicken genomic regions are complex, with numerous tandem duplications.

Table 1. Results of the fluorescence *in situ* hybridisation (FISH) localisation for the nine gene groups (GrA to GrI), identifying nine chromosomal regions with *PHF7*-like genes on *G. gallus* chromosomes with the selected BAC.

Gene group	**Selected BAC**	**FISH localisation**	**Measures**
A	WAG-038I15	GGA19	With WAG-062P02
B	WAG-119F08	GGAZ q12-13	Flcen: 38.8 ± 5.3
C	WAG-041O23	GGAZ p12-21	Flcen: 50.9 ± 7.0
D	WAG-023G20	GGA1 p26	Flpter: 3 ± 1.8
E	WAG-038A02	GGA1 p26	No measure
F	WAG-037B06	GGA2 q11-21	Flpter: 51.4 ± 2.5
G	WAG-035C03	GGAZ p21-22 GGA12	Flcen GGA Z: 34.7 ± 5.9 With WAG-033L02
H	WAG-034E06	GGA2 p11-12	Flpter: 25.6 ± 1.7
I	WAG-119J04	GGA1 q35-41	Flpter: 90.5 ± 2.2

FLpter refers to the fractional length of the chromosome from the telomere of the p arm (%). *FLcen* refers to the fractional length of the chromosome from the centromere (%).

Table 2. Details of the sequenced bacterial artificial chromosomes (BAC A to I).

BAC	Size (bp)	Total number of genes	Number of *PHF7*-like genes	Numbering of *PHF7*-like genes
A	103,706	19	3	13, 15, 17
B	122,743	23	9	5, 7, 8, 10, 12, 14, 16, 18, 20
C	147,352	9	1	6
D	147,939	9	2	1, 3
E	88,262	8	8	1, 2, 3, 4, 5, 6, 7, 8
F	96,796	4	1	1
G	86,864	10	4	1, 2, 3, 4
H	154,838	6	1	5
I	177,594	28	10	10, 12, 14, 16, 18, 20, 22, 24, 26, 28
Total	-	*116*	*39*	-

For each sequenced BAC A to I, the total gene number was identified by FGENESH in Softberry. For the details of each BAC annotation (positions of predicted genes) see Supplemental Table 4. The number and numbering* of *PHF7*-like genes is based on BLASTP results (as described in Supplemental Table 5).

bp= base pair

*The numbering is as described on the BAC maps in Fig 2 (BAC I) and Supplemental Fig 3 (BAC A to H).

DISCUSSION

In the present study, using FISH we identified 10 loci with *PHF7*-like genes in *G. gallus*, dispatched on five chromosomes: **Chr1** with three distinct loci, **ChrZ** with three distinct loci, **Chr2** with two distinct loci and **Chr12** and **Chr19** each with a unique locus. We also observed nine of these loci by using *in silico* PCR, and some of them had also been referenced in former EnsEMBL version 88 (then 'disappeared' in the genome annotation, as shown in Supplemental Table 2). According to our phylogenic tree, the first duplication concerned the locus on Chr 19 (BAC A) into 2 ancestors, one at the root of BAC F/G (Chr 2 and 12, respectively) and the other at the root of BAC H/D/E/C/B/I. In this latter group, BAC D and E (Chr1) have a common ancestor, and BAC B, C (both on Chr Z) and I (Chr 1) have a common ancestor, these two ancestors sharing a common older ancestor with BAC H (Chr 2).

Currently in EnsEMBL (release 105, [December 2021]), only four loci are described corresponding to four distinct genes (Supplemental Table 2, columns K and L), with one being on Chr12 with coordinates included in BAC G, and three on Chr1, with two genes included in BAC I and one (ENSG00000048616) on a different locus that we have not characterised. Thus, it is likely that we missed this locus and that the number of *PHF7*-like genes whose existence we demonstrated in the present study is underestimated. Another point of putative under-estimation is that we did not target any BAC corresponding to the second hybridisation of BAC G on ChrZ (with a different locus from BAC B and BAC C also on ChrZ), which may contain several other *PHF7*-like genes. Moreover, one can also imagine that there are *PHF7*-like genes present in genomic regions present in the upstream 5′ extremity and the downstream 3′ extremity of the sequenced BAC, especially for BAC that contain *PHF7*-like genes at their extremities, as is the case for BAC I and BAC E. A limit of our approach is that it does not allow sequencing BAC that were not detected by the screen with the initial primers (which are dependent on sequences available in the databases). Nevertheless, the chicken genome is relatively 'young' compared with the well-sequenced mouse genome (in which we have characterised massive duplications of reproductive genes [3-6, 10]) or the human genome. The quality of the chicken genome assembly may be optimised further in the future. Moreover, the chicken genome exhibits microchromosomes that are very difficult to sequence (GC-rich sequences). In the present study, we used the PacBio method to sequence the BAC [20]. The advantage of using PacBio's Circular Consensus Sequence (CCS) method is that it provides very high-fidelity, quality reads and allows obtaining a unique contig for each BAC clone. The technique allows correcting the sequence reads – the greater the number of repeated passes, the higher the Phred quality value

(QV) – and it allows considerably reducing polymorphisms that could be due to the technology itself (polymerase bias). Coupled with the evolution of PacBio chemistry, which allows reads between 15 and 20 kb (compared with ~7–8 kb with the old chemistry), it presents the great advantage to go beyond the repeated zones (whose average size is 10 kb) and to obtain a good quality assembly [21-23].

The technical strength of PacBio sequencing lies in new tools allowing for greater sequencing depth, thus better alignment and high-quality assembly. Thus, complex genomic regions that are difficult to access with more classical sequencing/assembly methods are now easier to access. Coupled with FISH to target specific genomic regions of the chicken bank of BAC available in our lab, it is very efficient to characterise a massive duplication family in the chicken genome, as we have done for *PHF7* in the present study.

Another source of difficulty and thus variability in results is the annotation process. High duplication is relatively rare, and thus the more recent EnsEMBL versions have probably simplified annotation with automatic algorithmic processes that eliminate massive duplications in the sequence, aiming to avoid putative false-positive gene redundancies. We have already observed this phenomenon with the expanding FFAR2 family: a version of EnsEMBL described a family of 23 paralogues, then in the following versions (including the current) only one *FFAR2* gene was present, whereas we experimentally found that the chicken genome contains 22 (± 2) paralogues [10].

Concerning our 'own' *ab initio* annotation of BAC sequences, we had the choice between three classical annotation programs available: Augustus, GENESCAN and Softberry, the last one based on FGENESH program that is the subject of a number of publications (as reviewed previously [24]). One inconvenience of GENESCAN is that no species can be targeted – only the vertebrate class. A convenient advantage of Softberry compared with the other two programs is the availability of a large amount of data – in particular, we retrieved each exon sequence so that we could align them. Annotation of predicted genes was almost identical for the three programs, with few differences. One concerns gene 1 in BAC F: Augustus and GENSCAN instead predicted three and two genes, respectively (the long gene 1 is split into three and two smaller genes, that are also *PHF7*-like genes). Subtleties in the annotation process/program may explain such variations [25], which may be a further source of underestimation of *PHF7*-like genes in the present results. Of note, all three programs predicted the strange long gene 5 in BAC H exhibiting 70 exons, corresponding to a repetition of seven *PHF7* 'core-structures'. We performed several trials by RT-PCR dedicated to long RNA, but we were unable to find a trace of this long mRNA (10 kb) that could correspond to gene 5 in BAC H. Moreover, we have our own NGS data from chicken testis (data not published but deposited in GEO[2]) but could not find long reads matching BAC H. Thus, it is still unclear whether such large mRNA (and its corresponding protein) exists or is an artefact and a consequence of imperfect annotation with existing tools. According to their exon structures (see suppl Fig 4), these long genes may be split in smaller "conventional" *PHF7*-like genes with 8-10 exons, and would thus correspond to: 7 genes in BAC H (instead of one 70-exons gene), 6 genes in BAC E (instead of two 30-exons genes) and 3 genes in BAC F (instead of one 28-exons gene). Thus, instead of 4 long genes, there may be 12 smaller *PHF7*-like genes (for a putative total of 52 *PHF7*-like genes instead of 39).

PHF7 protein is expressed in male germ cells during spermiogenesis and involved in histone-to-protamine exchange. In *Drosophila melanogaster*, deletion mutants of *Phf7* have demonstrated the important role of this gene for male fertility [16, 26]. Male infertility in mice with *Phf7* deletion is due to aberrant histone retention and impaired protamine replacement in elongated spermatids [27]. In the chicken, as in the rat and human, *PHF7* mRNA expression is much higher, if not even exclusive, in the testis compared with other tissues [14, 18]. In this species, we observed an increase in the mRNA level with the animal's age (data not shown), suggesting germ cell expression in the chicken as in other species. Additional studies would be needed to better characterise the protein expression and function in the chicken testis. At the evolutionary level, a previous study established that one copy of the gene is present in the *Drosophila* and mammalian genomes, whereas several copies (but not characterised/counted) are present on two loci in the chicken genome [18]. Moreover, these authors showed that *Phf7* has a common ancestor with *G2e3* (G2/M-phase specific E3 ubiquitin protein ligase). Both genes arose from a duplication before the divergence of vertebrates, and non-vertebrates have only one gene of this family. These two proteins possess three zinc fingers (PHD domains and RING fingers, respectively, for *Phf7* and *G2e3*) in their N-terminus. In their study, Wang et al. [18] showed that *G2e3* is present in all metazoan genomes whereas *Phf7* is absent in fish and reptile genomes. Currently, however, there are reptile orthologues of the mouse *Phf7* gene that can be found in EnsEMBL release 105 (a unique orthologue in each of the Goodes thorn scrub tortoise, the painted turtle, the Abingdon Island giant tortoise; two in the three-toed box turtle). Because there are numerous predicted duplications of *PHF7* in other birds (EnsEMBL release 105: turkey, 24; Japanese quail, 9; zebra finch, 2; collared flycatcher, 2; duck, 5), we

[2] https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE133401 (available in 2023).

hypothesise that the massive *PHF7* duplication is restricted to birds (at least until better annotation is available in reptiles), with no hypothesis of their biological sense (if any).

In the chicken genome, we previously characterised the *FFAR2* [10] gene massive duplication (22 paralogues, whereas a unique gene is present in the mouse and human) and exhibiting a high testicular level of expression. In our more recent study, we identified, in addition to *PHF7*, two other chicken testicular family genes with many paralogues (whereas a unique gene in the mouse): *SUN3* (11 paralogues) and *SPAG4*/*SUN5* (20 paralogues) [14]. As for *PHF7* and for reasons discussed above, the presence of these numerous paralogous genes fluctuates according to databases and their successive versions (for example, there is a unique *FFAR2* gene in the chicken genome according to current EnsEMBL version 105). Studies similar to the one we have conducted here for *PHF7* would be of interest to characterise such families; however, such work is time-consuming and expensive. Better annotation of the *G. gallus* genome (Galgal7) may improve identification of massively duplicated regions. Another interesting point would be to improve annotation in other bird species, as well as in other sauropsids, aiming to conduct evolutionary studies to date massive duplications.

From a functional point of view, the significance of the presence of several, and sometimes many, paralogues in animal genomes is still unclear. In the mouse, when paralogues encode proteins with similar sequences and tissue expression, as is this the case for *PHF7* and other genes mentioned above, they are quite often, but not always, able to compensate for the loss (in mutants) of their paralogues. Indeed, there are cases of dispensability of paralogous genes and there are also cases of non-redundancy of paralogues [28]. For example, for *Nlrp5* (or Mater) and its paralogue *Nlrp4e* characterised in the lab [3, 6], individual invalidation of each of them leads in both cases to a drastic phenotype of sterility with early embryonic death [29, 30]. Guschanski et al. [31] showed that, during vertebrate evolution, the contribution of paralogues to specific organ functions differs according to the organ, with paralogues expressed in young testis putatively involved in lineage-specific biology consistently with their reproductive function.

Overall, we have characterised a new expanding germ cell–specific gene family in the chicken genome. The PHF7-like proteins and genes exhibit a strong level of similarity (as stated by rakes in their phylogenic tree) that may be explain by gene conversion process; this hypothesis would need future specific studies. Also, its functional and evolutionary significance (as for other germinal specific gene families as Oogenesin or Nlrp5 in the mouse) remain to be investigated further, given the lack/absence of information in the genome database due to the technical problems mentioned above. We need to clarify and illuminate ‘the dark side’ of the chicken genome, such regions with recent duplications still being the blind spot of genomic sequencing programmes.

LEGENDS

Figure 1.

(A) Fluorescence *in situ* hybridisation (FISH) localisations (white arrows) of bacterial artificial chromosome (BAC) clones (red signals) screened for groups of genes (GrA to GrI) based on the international standard of the chicken karyotype [32]. Green signals are for microchromosome detection (GrA and GrG). GrA (A), WAG-038I15 (red) together with WAG-062P02 (green) on GGA19; GrB (B), WAG-119F08 on GGAZq12-13; GrC (C), WAG-041O23 on GGAZp12-21; GrD (D), WAG-023G20 on GGA1p26; GrE (E), WAG-038A02 on GGA1p26; GrF (F), WAG-037B06 on GGA2q11-21; GrG (G), WAG-035C03 on GGAZp21-22 and together with WAG-033L02 (green) on GGA12; GrH (H), WAG-034E06 on GGA2p11-12; GrI (I), WAG-119J04 on GGA1q35-41.

(B) Fluorescence *in situ* hybridisation (FISH) localisations of the gene groups (GrA to GrI) obtained in Fig 1A drawn on the chicken standard idiograms [32] from the measures summarised in Table 1.

Figure 2.

BAC I gene map obtained after annotation by FGENESH and BLASTN/BLASTP. The grey box represents the BAC I with its first nucleotide in position 1. The coloured boxes represent the 22 genes predicted by FGENESH (with their relative coordinates on the BAC); the colours of the text ‘Gen#’ refer to BLASTP results described in Supplemental Table 5, with families in colour (uncoloured genes are unique): Yellow = PHF7 family; red = translation initiation factor IF-2-like isoform X2 family; pink = reverse transcriptase family; green = centrosomal protein of 126 kDa family. Other BAC maps are in Supplemental Fig 3.

Figure 3.

Phylogenetic tree of the 39 PHF7-like proteins (sequences are in Supplemental Table 6). See the Materials and methods for details on its construction. The bootstrap values are in red.

MATERIALS AND METHODS

In silico and genomic PCR to characterize chicken *PHF7* genes

We used *PHF7* primers (nine pairs) listed in Supplemental Table 3 (column B) that we had designed in our previous work [14] (to study mRNA expression). These primers were designed using NCBI "primer-blast" tool. It was not possible to design a unique primer pair for the 63 EnsEMBL genes, but a pair was obtained for each of the 9 subtrees/groups (suppl Fig 1 and suppl Table 3), allowing to cover all sequences of chicken paralogs. In the present study, we used them to perform *in silico* genomic PCR with the tool available at https://genome.ucsc.edu/cgi-bin/hgPcr. Genomic PCR (30 cycles: 95°C for 10 s, 60°C for 10 s and 72°C for 30 s), with chicken DNA extracted from the blood of Leghorn chicken (pool of three animals; kindly provided by Amélie Juanchich, BOA INRAE F-37380 Nouzilly) allowed us to verify their specificity on agarose gel (single amplicons at their theorical sizes), before being used for probe production as described below.

FISH

For each of the nine groups (GrA to GrI), BAC clones were selected from the Wageningen chicken (White Leghorn breed) library by two-dimensional PCR screening of super-pools and pools arranged in microplates as described by Crooijmans et al. [33]. PCR amplifications were carried out for each group by using the primers listed in Supplemental Table 3 (column B) as follows: 35 cycles with denaturation at 95°C for 30 s, specific annealing at 60°C for 30 s and elongation at 72°C for 30 s. Each 20 µl reaction contained 2 mM $MgCl_2$, 0.2 mM dNTPs, 0.5 µM primers and 0.625 units Taq polymerase (Go Tad Flexi DNA polymerase Promega™ M3005). The reactions were run on an Applied Biosystems™ 2720 Thermal Cycler.

After isolating a single colony on a Petri dish to avoid any risk of contamination, BAC clones were grown in 25 ml of LB medium with 34 µg/ml chloramphenicol. The DNA was extracted based on alkaline lysis using the Qiagen Plasmid Midi Kit. The presence of each group of genes in the corresponding BAC clone was checked by PCR as described previously. PCR products were sequenced by using the Sanger technique on the Get-Plage Genotoul Platform (GeT-PlaGe INRAE Auzeville F-31326 Castanet-Tolosan Cedex France) to confirm the gene identities. The sequenced were visualised with Chromas software and aligned by Blast.

FISH was carried out on metaphase spreads obtained from fibroblast cultures of 7-day-old chicken and duck embryos, arrested with 0.05 µg/ml colcemid (Sigma). After a 10 min hypotonic treatment (1:5 foetal calf serum hypotonic solution mixed equal parts with 0.075 M

KCl), the cell suspension was fixed overnight in a 3:1 ratio of ethanol to acetic acid solution and stored at -20°C until spreading.

The single-colour FISH protocol is based on Yerle et al. [34]. Briefly, 150 ng of DNA of each BAC clone was biotin labelled (biotin 16-dUTP) by random priming using the Bioprim Kit (Invitrogen). The probes were purified using MicroSpin G-50 columns (GE Healthcare Life sciences) to remove the non-incorporated nucleotides. Probes were ethanol precipitated and resuspended in 50% formamide hybridisation buffer. After denaturation of probes (7 min at 100°C) and chromosomes (2 min at 72°C in 70% formamide), slides were hybridised *in situ* for 17 h at 37°C in the presence of 5 µg chicken cot1 competitor DNA on a humid plate (Dako Hybridizer). After hybridisation, slides were washed 2 × 30 min in 2X SSC then 4 min at 73°C in 0.4X SSC. The biotin was detected with Alexa568-Streptavidin (from Invitrogen).

For group A (WAG-038I15) and G (WAG-035C03), expected to be located on a microchromosome pair, the corresponding BAC were co-hybridised with specific FISH markers (WAG-062P02 for GGA19 and WAG-033L02 for GGA12) used as references to identify precisely the microchromosome pairs involved [35-37]. Two-colour FISH was performed according to Trask et al. (1991) [38]. One probe was labelled with digoxigenin (digoxigenin-11-dUTP, Roche) and the other with biotin (biotin 16-dUTP) using the BioPrime Kit (Invitrogen). The two labelled probes were ethanol precipitated together before hybridisation. The biotin-labelled probe was detected with Alexa568-streptavidin and the digoxigenin labelled probe was detected with Alexa488-anti-digoxigenin (from Invitrogen).

Chromosomes were counterstained with DAPI in antifade solution (Vectashield with DAPI, Vector Laboratories-H-1200). The hybridised metaphases were screened with a Zeiss fluorescence microscope; a minimum of 20 spreads were analysed for each experiment. Spot-bearing metaphases were captured and analysed with a cooled CCD camera using Cytovision software (Leica Biosystem).

We defined the precise localisations for macrochromosomes by the fractional length from the p arm telomere (Flpter) after measurement of 10 chromosomes (Cytovision software), except for ChrZ (GGAZ), for which we used the fractional length from the centromere (Flcen) because this chromosome is difficult to orientate. We used Flpter and Flcen to determine the FISH localisation on the G-banded chicken standard karyotype as shown on Fig 1B [32].

BAC clone sequencing: PacBio library preparation, sequencing and data assembly

Individual BAC clone DNA was extracted by using the Nucleobond Xtra Midi Kit (Macherey-Nagel). Two micrograms of each sample was used to construct a multiplexed SMRTbell® library by the INRAE-CNRGV. We followed the PacBio recommendations for Multiplexed Microbial Library preparation (PN 101-696-100) with some adjustments by using the SMRTbell Express Prep kit v2.0 (Pacific Biosciences, Menlo Park, CA, USA). The first enzymatic steps consist of removing single-stranded overhangs, repairing any DNA damage and polishing the ends of the double-stranded fragments and tailing with an A-overhang. Ligation with specific barcoded hairpin T-overhang adapters to both ends of the targeted double-stranded DNA (dsDNA) molecule creates a closed, single-stranded circular DNA. Each individual sample was treated with nuclease by using SMRTbell Enzyme Clean-up kit (Pacific Biosciences). The Blue-Pippin size-selection system (Sage Science, Beverly, MA, USA) was used to remove fragments < 15 kb from pooled sample previously purified with 0.45X AMPure PB beads (Pacific Biosciences). The size and concentration of the final library were assessed using the FemtoPulse system and the Qubit Fluorometer and Qubit dsDNA HS reagents Assay kit (Thermo Fisher Scientific, Waltham, MA, USA), respectively.

Sequencing primer v2 and Sequel DNA Polymerase 2.0 were annealed and bound, respectively, to the SMRTbell library. The library was loaded onto one SMRTcell at an on-plate concentration between 50 and 85 pM by using a diffusion loading. Sequencing was performed on the Sequel II system with a run movie time of 30 h with 120 min pre-extension step and Software v9.0 (PacBio) by Gentyane Genomic Platform (INRAE-Clermont-Ferrand, France).

We corrected the PacBio raw reads by using SMRTLink_v9.0.0 with eight passes, then demultiplexed the data. We identified residual *Escherichia coli* reads by using BLAST+ 2.10.0 and removed them by using Seqfilter. We filtered the HiFi reads by identifying the vector sequences using cross_match and removed them with custom Perl scripts. We filtered HiFi reads < 15 kb by using Seqfilter, and then subsampled with SeqKit to obtain an estimated average assembly depth of 50X. We assembled the reads with hifiasm-0.12. To validate the result, we checked the length of the obtained contig and mapped BAC end sequences with the extremities on the assembly using BLAST+ 2.10.0. We remapped HiFi reads to the assembly and obtained the depth with samtools-1.8.

***In silico* analyses (except *in silico* PCR)**

Orthology link

We obtained *in silico* data concerning *PHF7* orthologues in birds (and other species) from EnsEMBL https://www.ensembl.org/ (from version 88 [March 2017] to the current version 105 [December 2021]).

Gene annotation

For gene structure prediction of the BAC nucleotide sequences, we performed *ab initio* annotation by using the commonly used FGENESH program on the Softberry site [39], which relies on hidden Markov model (HMM) statistical models to identify promoters, coding or noncoding regions, and intron–exon junctions (available at http://www.softberry.com/) [24]. When needed, we consulted two other classical programs: Augustus (http://bioinf.uni-greifswald.de/augustus/submission.php) and GENESCAN (http://hollywood.mit.edu/GENSCAN.html).

Similarity with sequences in databases using NCBI BLAST

We searched for similarity and/or functional homology by using the BLAST tool of NCBI (https://blast.ncbi.nlm.nih.gov/Blast.cgi). We performed BLASTN with the nucleotide sequence of each BAC on the galGal7 genome (GCRg7w, white Leghorn layer, NCBI 106). We performed BLASTP with the 116 peptides (sequences in Supplemental Table 6) obtained with FGENESH. The criteria of the BLASTP interrogation were 'by default' or, when no results were obtained, the threshold was upgraded (we arbitrarily chose 1000; in this case, E-value may be high, i.e. > 0).

Phylogeny and sequence alignment and similarity

We constructed the phylogenetic tree from the PHF7-like protein sequences with http://www.phylogeny.fr/alacarte.cgi (MUSCLE for Multiple Alignment; Gblocks for Alignment curation; construction of phylogenetic tree with PhyML; visualisation of phylogenetic tree with TreeDyn). The bootstrap values were estimated with 1000 replications and the tree was rooted using midpoint rooting method. We created a representation of the PHF7 sequence alignment with the program ESPript (Easy Sequencing in PostScript, available at https://espript.ibcp.fr/ESPript/ESPript/) [40], which displays sequence similarities from aligned sequences.

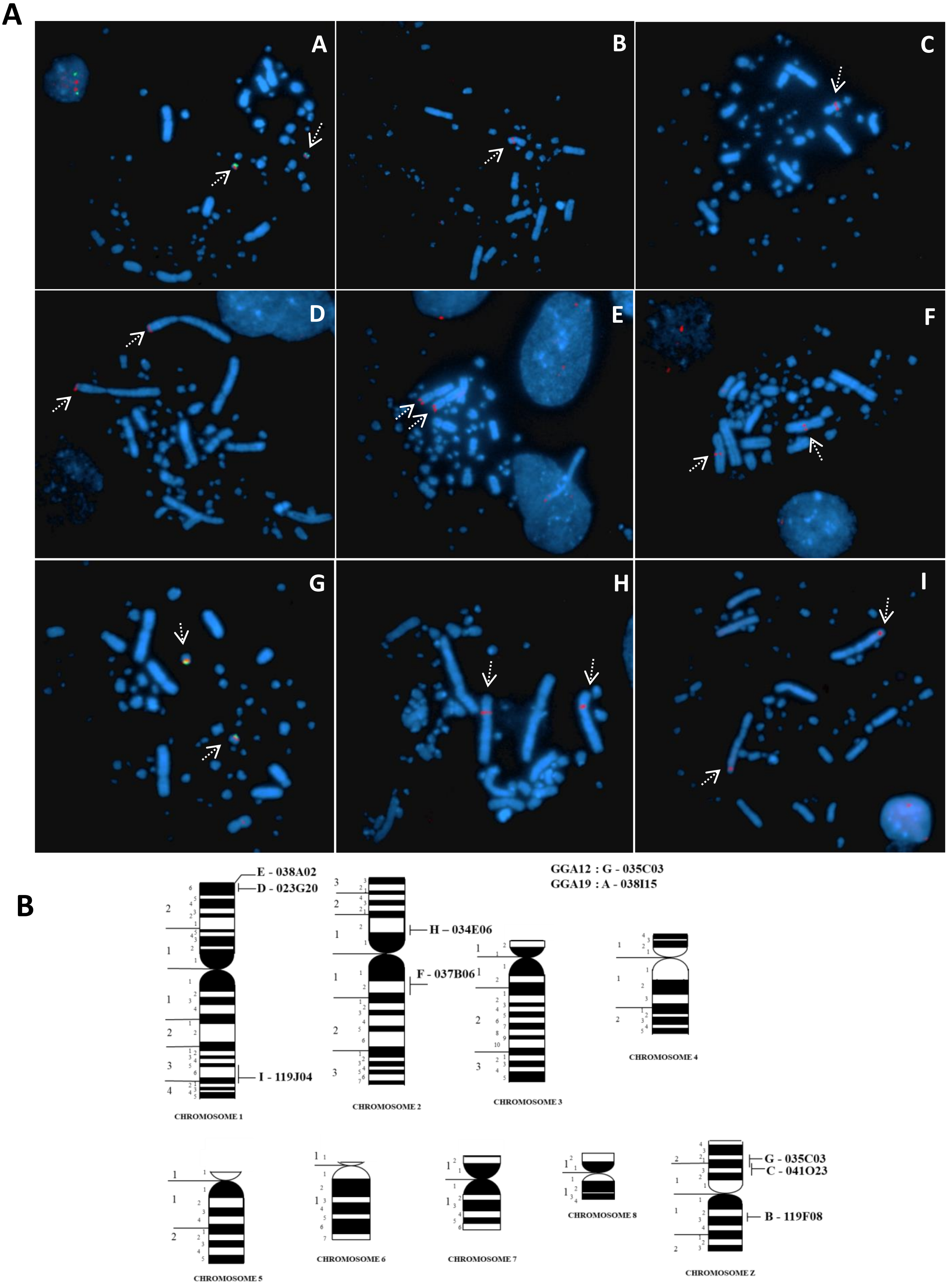
Acknowledgments

SF and PM are very grateful to Charlotte Lécureuil for her careful re-reading of the manuscript. We thank Amélie Juanchich (INRAE BOA Nouzilly) for providing chicken DNA. The manuscript was proofread and corrected by Proof-Reading-Service.com.

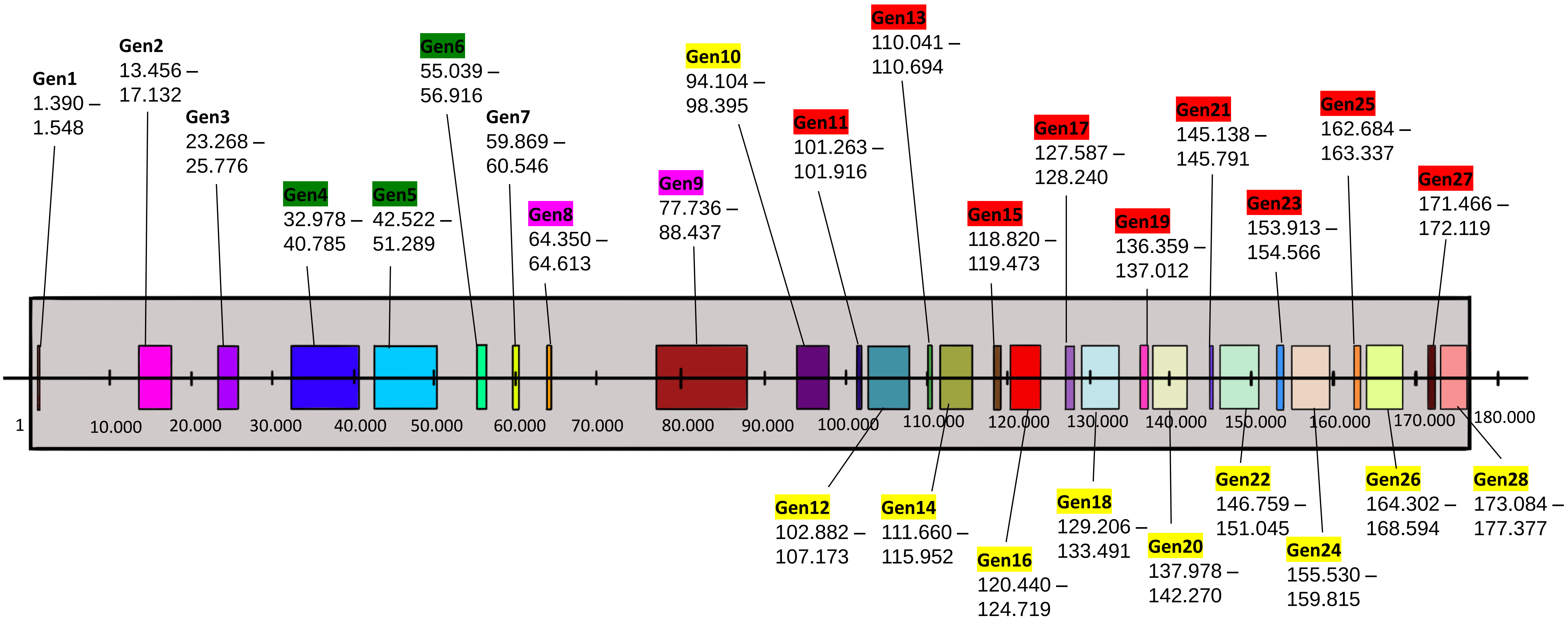
References

1. Demuth, J.P. and M.W. Hahn, *The life and death of gene families.* Bioessays, 2009. **31**(1): p. 29-39.
2. Taylor, J.S. and J. Raes, *Duplication and divergence: the evolution of new genes and old ideas.* Annu Rev Genet, 2004. **38**: p. 615-43.
3. Dade, S., et al., *Identification of a new expanding family of genes characterized by atypical LRR domains. Localization of a cluster preferentially expressed in oocyte.* FEBS Lett, 2003. **555**(3): p. 533-8.
4. Dade, S., et al., *In silico identification and structural features of six new genes similar to MATER specifically expressed in the oocyte.* Biochem Biophys Res Commun, 2004. **324**(2): p. 547-53.
5. Paillisson, A., et al., *Identification, characterization and metagenome analysis of oocyte-specific genes organized in clusters in the mouse genome.* BMC Genomics, 2005. **6**: p. 76.
6. Tian, X., G. Pascal, and P. Monget, *Evolution and functional divergence of NLRP genes in mammalian reproductive systems.* BMC Evol Biol, 2009. **9**: p. 202.
7. Lee, M.O., et al., *Duplication of chicken defensin7 gene generated by gene conversion and homologous recombination.* Proc Natl Acad Sci U S A, 2016. **113**(48): p. 13815-13820.
8. Temperley, N.D., et al., *Evolution of the chicken Toll-like receptor gene family: a story of gene gain and gene loss.* BMC Genomics, 2008. **9**: p. 62.
9. Zhang, T., et al., *Genomic organization of the chicken TCRβ locus originated by duplication of a Vβ segment combined with a trypsinogen gene.* Vet Immunol Immunopathol, 2020. **219**: p. 109974.
10. Meslin, C., et al., *Expanding Duplication of Free Fatty Acid Receptor-2 (GPR43) Genes in the Chicken Genome.* Genome Biol Evol, 2015. **7**(5): p. 1332-48.
11. Bhasin, S., C. Mallidis, and K. Ma, *The genetic basis of infertility in men.* Baillieres Best Pract Res Clin Endocrinol Metab, 2000. **14**(3): p. 363-88.
12. Bonilla, E. and E.Y. Xu, *Identification and characterization of novel mammalian spermatogenic genes conserved from fly to human.* Mol Hum Reprod, 2008. **14**(3): p. 137-42.
13. Rodgers-Melnick, E.B. and R.K. Naz, *Male-biased genes of Drosophila melanogaster that are conserved in mammalian testis.* Front Biosci (Elite Ed), 2010. **2**: p. 841-8.
14. Fouchecourt, S., et al., *An evolutionary approach to recover genes predominantly expressed in the testes of the zebrafish, chicken and mouse.* BMC Evol Biol, 2019. **19**(1): p. 137. doi: 10.1186/s12862-019-1462-8.
15. Kim, C.R., et al., *PHF7 Modulates BRDT Stability and Histone-to-Protamine Exchange during Spermiogenesis.* Cell Rep, 2020. **32**(4): p. 107950.
16. Yang, S.Y., E.M. Baxter, and M. Van Doren, *Phf7 controls male sex determination in the Drosophila germline.* Dev Cell, 2012. **22**(5): p. 1041-51.
17. Merkin, J., et al., *Evolutionary dynamics of gene and isoform regulation in Mammalian tissues.* Science, 2012. **338**(6114): p. 1593-9.
18. Wang, X.R., et al., *Evidence for parallel evolution of a gene involved in the regulation of spermatogenesis.* Proc Biol Sci, 2017. **284**(1855). doi: 10.1098/rspb.2017.0324.
19. English, A.C., et al., *Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology.* PLoS One, 2012. **7**(11): p. e47768.
20. Athanasopoulou, K., et al., *Third-Generation Sequencing: The Spearhead towards the Radical Transformation of Modern Genomics.* Life (Basel), 2021. **12**(1). doi: 10.3390/life12010030.

21. Hon, T., et al., *Highly accurate long-read HiFi sequencing data for five complex genomes.* Sci Data, 2020. **7**(1): p. 399.
22. Murigneux, V., et al., *Comparison of long-read methods for sequencing and assembly of a plant genome.* Gigascience, 2020. **9**(12).
23. Wenger, A.M., et al., *Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome.* Nat Biotechnol, 2019. **37**(10): p. 1155-1162.
24. Ejigu, G.F. and J. Jung, *Review on the Computational Genome Annotation of Sequences Obtained by Next-Generation Sequencing.* Biology (Basel), 2020. **9**(9). doi: 10.3390/biology9090295.
25. Mudge, J.M. and J. Harrow, *The state of play in higher eukaryote gene annotation.* Nat Rev Genet, 2016. **17**(12): p. 758-772.
26. Yang, S.Y., et al., *Control of a Novel Spermatocyte-Promoting Factor by the Male Germline Sex Determination Factor PHF7 of Drosophila melanogaster.* Genetics, 2017. **206**(4): p. 1939-1949.
27. Wang, X., et al., *PHF7 is a novel histone H2A E3 ligase prior to histone-to-protamine exchange during spermiogenesis.* Development, 2019. **146**(13).
28. Roth, C., et al., *Evolution after gene duplication: models, mechanisms, sequences, systems, and organisms.* J Exp Zool B Mol Dev Evol, 2007. **308**(1): p. 58-73.
29. Chang, B.H., et al., *Developmental expression and possible functional roles of mouse Nlrp4e in preimplantation embryos.* In Vitro Cell Dev Biol Anim, 2013. **49**(7): p. 548-53.
30. Tong, Z.B., et al., *Mater, a maternal effect gene required for early embryonic development in mice.* Nat Genet, 2000. **26**(3): p. 267-8.
31. Guschanski, K., M. Warnefors, and H. Kaessmann, *The evolution of duplicate gene expression in mammalian organs.* Genome Res, 2017. **27**(9): p. 1461-1474.
32. Ladjali-Mohammedi, K., et al., *International system for standardized avian karyotypes (ISSAK): standardized banded karyotypes of the domestic fowl (Gallus domesticus).* Cytogenet Cell Genet, 1999. **86**(3-4): p. 271-6.
33. Crooijmans, R.P., et al., *Two-dimensional screening of the Wageningen chicken BAC library.* Mamm Genome, 2000. **11**(5): p. 360-3.
34. Yerle, M., et al., *Localization of the pig luteinizing hormone/choriogonadotropin receptor gene (LHCGR) by radioactive and nonradioactive in situ hybridization.* Cytogenet Cell Genet, 1992. **59**(1): p. 48-51.
35. Fillon, V., et al., *Identification of 16 chicken microchromosomes by molecular markers using two-colour fluorescence in situ hybridization (FISH).* Chromosome Res, 1998. **6**(4): p. 307-13.
36. Masabanda, J.S., et al., *Molecular cytogenetic definition of the chicken genome: the first complete avian karyotype.* Genetics, 2004. **166**(3): p. 1367-73.
37. Douaud, M., et al., *Addition of the microchromosome GGA25 to the chicken genome sequence assembly through radiation hybrid and genetic mapping.* BMC Genomics, 2008. **9**: p. 129.
38. Trask, B.J., et al., *Mapping of human chromosome Xq28 by two-color fluorescence in situ hybridization of DNA sequences to interphase cell nuclei.* Am J Hum Genet, 1991. **48**(1): p. 1-15.
39. Solovyev, V., et al., *Automatic annotation of eukaryotic genes, pseudogenes and promoters.* Genome Biol, 2006. **7 Suppl 1**(Suppl 1): p. S10.1-12.
40. Gouet, P., X. Robert, and E. Courcelle, *ESPript/ENDscript: Extracting and rendering sequence and 3D information from atomic structures of proteins.* Nucleic Acids Res, 2003. **31**(13): p. 3320-3.


A
A
B
C
D
E
F
G
H
I
B
E - 038A02
D - 023G20
I - 119J04
CHROMOSOME 1
H – 034E06
F - 037B06
CHROMOSOME 2
CHROMOSOME 3
GGA12 : G - 035C03
GGA19 : A - 038I15
CHROMOSOME 4
CHROMOSOME 5
CHROMOSOME 6
CHROMOSOME 7
CHROMOSOME 8
G - 035C03
C - 041O23
B - 119F08
CHROMOSOME Z

BAC I : 177.594 bp
Blastn result: Chromosome 1 with following coordinates
NCBI version 106 (GRCg7w): 182.685.165 - 182.769.800



Gen1 1.390 – 1.548

Gen2 13.456 – 17.132

Gen3 23.268 – 25.776

Gen4 32.978 – 40.785

Gen5 42.522 – 51.289

Gen6 55.039 – 56.916

Gen7 59.869 – 60.546

Gen8 64.350 – 64.613

Gen9 77.736 – 88.437

Gen10 94.104 – 98.395

Gen11 101.263 – 101.916

Gen12 102.882 – 107.173

Gen13 110.041 – 110.694

Gen14 111.660 – 115.952

Gen15 118.820 – 119.473

Gen16 120.440 – 124.719

Gen17 127.587 – 128.240

Gen18 129.206 – 133.491

Gen19 136.359 – 137.012

Gen20 137.978 – 142.270

Gen21 145.138 – 145.791

Gen22 146.759 – 151.045

Gen23 153.913 – 154.566

Gen24 155.530 – 159.815

Gen25 162.684 – 163.337

Gen26 164.302 – 168.594

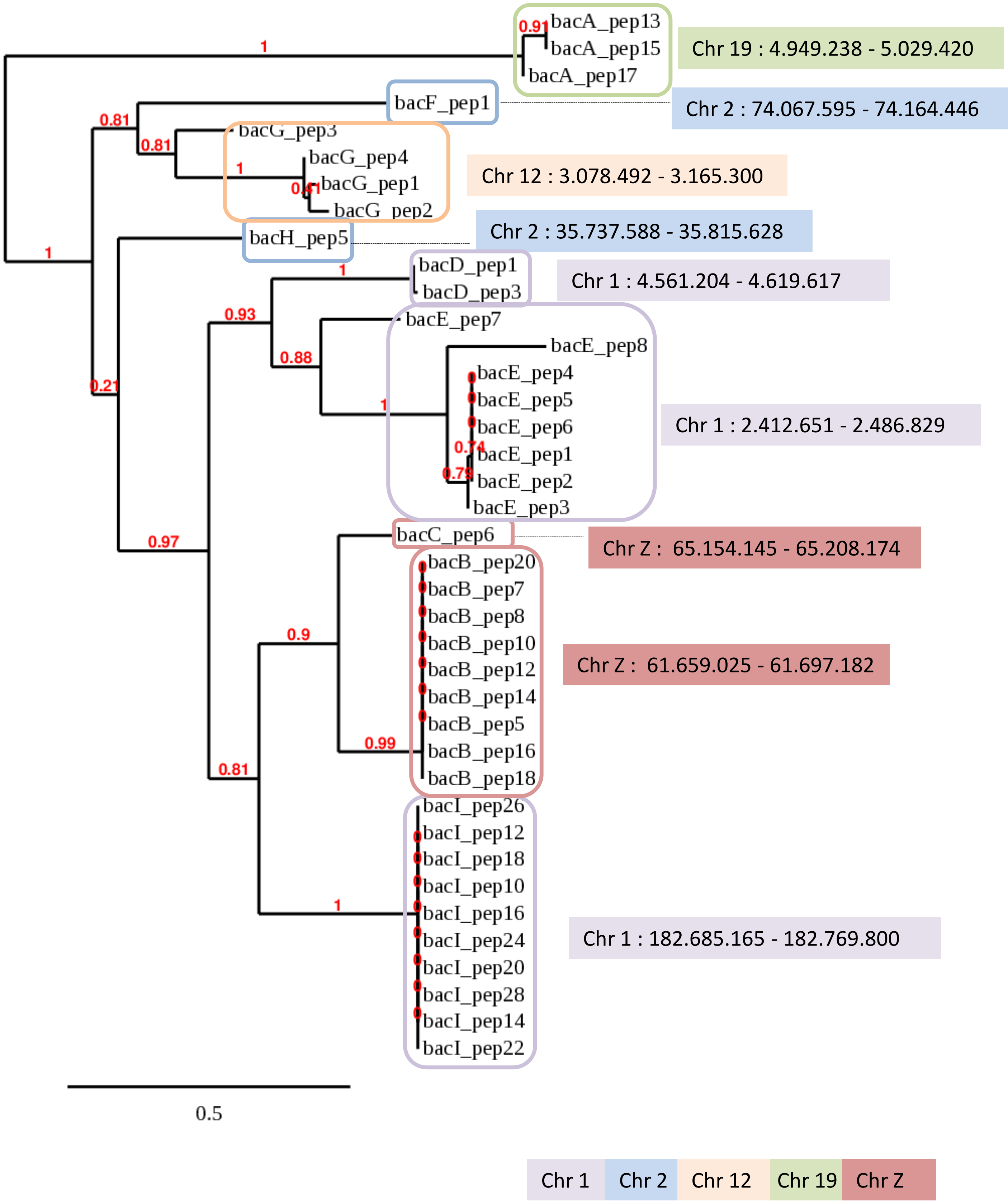
Gen27 171.466 – 172.119

Gen28 173.084 – 177.377

Gen4/5/6: centrosomal protein of 126 kDa

Gen10/12/14//16/18/20/22/24/26/28: *PHF7*-like

Gen11/13/15/17/19/21/23/25/27: translation initiation factor family


1
0.91
bacA_pep13
bacA_pep15
bacA_pep17
Chr 19 : 4.949.238 - 5.029.420
0.81
bacF_pep1
Chr 2 : 74.067.595 - 74.164.446
0.81
bacG_pep3
1
bacG_pep4
0.41
bacG_pep1
bacG_pep2
Chr 12 : 3.078.492 - 3.165.300
1
bacH_pep5
Chr 2 : 35.737.588 - 35.815.628
1
bacD_pep1
bacD_pep3
Chr 1 : 4.561.204 - 4.619.617
0.93
bacE_pep7
0.88
bacE_pep8
0.21
bacE_pep4
bacE_pep5
bacE_pep6
1
0.74
bacE_pep1
0.79
bacE_pep2
bacE_pep3
Chr 1 : 2.412.651 - 2.486.829
0.97
bacC_pep6
Chr Z : 65.154.145 - 65.208.174
bacB_pep20
bacB_pep7
bacB_pep8
0.9
bacB_pep10
bacB_pep12
Chr Z : 61.659.025 - 61.697.182
bacB_pep14
bacB_pep5
0.99
bacB_pep16
0.81
bacB_pep18
bacI_pep26
bacI_pep12
bacI_pep18
bacI_pep10
1
bacI_pep16
bacI_pep24
Chr 1 : 182.685.165 - 182.769.800
bacI_pep20
bacI_pep28
bacI_pep14
bacI_pep22
0.5
Chr 1
Chr 2
Chr 12
Chr 19
Chr Z